基于 CNN 和 LSTM 混合模型的人体跌倒行为研究 *

库向阳，苏学威

(西安科技大学 计算机科学与技术学院，西安 710054)

摘 要：视频监控中人体跌倒行为识别对于提升老年人护理质量，减少社会养老负担等方面有十分重要意义。传统模式识别方法依赖于人工选取的特征，智能化程度低，识别精度不高。深度学习模型泛化能力强，特征提取自动完成。但目前深度学习模型不能较好的把监控视频中跌倒行为的空间和时序特征有效结合起来。为此，提出基于CNN(convolutional neural network)和 LSTM(long-short term memory)混合模型的人体跌倒行为识别方法。该模型采用两层结构，将视频以每 5 帧为一组输入到网络中，CNN 提取视频序列的空间特征，LSTM 提取视频时间维度上的特征，最后使用 softmax 分类器进行识别。实验表明，该方法可以有效提高跌倒识别的准确率。

关键词：跌倒行为识别；卷积神经网络；长短期记忆网络；时间维度

中图分类号：TP391.41　　doi: 10.3969/j.issn.1001-3695.2018.06.0424

Research on human fall behavior using CNN and LSTM-based hybrid model

She Xiangyang, Su Xuewei

(*College of Computer Science & Technology*, *Xi'an University of Science & Technology*, *Xi'an* 710054, *China*)

Abstract: The detection of human fall behavior in video surveillance is of great significance for improving the quality of care for the elderly and reducing the burden of social pension. The traditional pattern recognition method relies on the characteristics of the manual selection, the degree of intelligence is low, and the recognition accuracy is not high. Deep learning model has strong generalization ability and can extract the feature automatically. However, the above models cannot effectively combine the spatial and temporal characteristics of the fall behavior in surveillance videos. To this end, This paper proposed a method which combinates CNN and LSTM (long-short term memory) models for the study of human fall behavior. The model adopted a two-layer structure and put the video into the network every 5 frames. The CNN extracted the spatial features of the video sequence. The LSTM extracted the features of the video in the time dimension. Finally, the softmax classifier obtained the classification result. Experiments show that this method can effectively improve the accuracy of fall recognition.

Key words: falling behavior recognition; convolutional neural network; long-short term memory; time dimension

0　引言

随着社会人口老龄化的发展，因为跌倒导致老人意外受伤或死亡的情况时有发生。准确高效地识别出监控视频中跌倒行为对于老年人的安全防护具有重要的现实意义[1]。许多学者在视频中跌倒行为识别方面做了大量研究，提出了一些识别方法。主要有两种：基于浅层传统模式识别方法和基于深度学习的分类模型。

a）在浅层传统模式识别方面，文献[2]人工提取人体轮廓外接矩形的宽高比、人体 Hu 矩特征、人体轮廓离心率、人体轴线角等多特征进行融合，采用 SVM 检测跌倒行为。文献[3]提取对跌倒行为敏感的时域及频域特征，利用奇异值分解方法降维和重构跌倒特征，采用 SVM 分类器检测跌倒行为。以上方法的识别率依赖事先人工提取的特征，一旦提取的特征不理想，跌倒行为识别的效果就会受到较大影响。

b）深度学习模型通过对数据多层建模获得视频数据的特征表示，避免了人工提取特征的繁琐，而且具有良好的泛化能力。文献[4]提出一种基于 CNN 深度学习人体行为识别方法，该方法由卷积神经网络进行局部特征分析，得到特征输出项进行分类。但是该方法仅仅得到了局部空间特征，丢失了时域特征。文献[5]中提到一种基于 LSTM 深度学习人体行为识别方法，对时间序列进行建模，对人体行为进行训练和识别。但该方法的不足之处在于其仅仅提取了视频数据的时序特征，而丢失了局部空间特征。

收稿日期：2018-06-27；**修回日期：**2018-08-25　**基金项目：**陕西省自然科学基金研究项目（2017JM6105）

作者简介：库向阳（1968-），男，教授，博士（后），主要研究方向为数据挖掘与智能信息处理、人工智能与模式识别、复杂系统建模与优化等（1535594191@qq.com）；苏学威（1993-），男，硕士研究生，主要研究方向为机器学习、图像处理.

为了更好地获取视频数据空间和时序特征，一些专家学者将 CNN 和 LSTM 结合起来，并成功应用于视频分类和视频描述方面。Ng[6]等将图像数据和光流数据分别通过 CNN，获取视频帧序列的空间信息，然后将 CNN 输出传入 LSTM，以挖掘它们之间的时序信息，最后通过 softmax 对视频类别进行预测。Venugopalan 等人[7]将短视频抽样为 16 帧图像序列并以此来代表整部视频，由 CNN 来提取特征，然后将此 16 帧特征做均值池化得到视频编码特征，然后利用 LSTM 解码生成视频描述信息。

在前人研究的基础上，我们提出基于 CNN 和 LSTM 的混合模型来对跌倒行为进行识别的方法，对视频数据集进行简单处理后，混合模型利用 CNN 滑动窗口和权值共享[8]来获得视频序列的局部空间特征并作为下一层的输入，利用 LSTM 的时序性获取视频数据的时间特征，将两者结合起来，充分利用了两者各自的优势。另外，由于深度学习可以自动提取行为特征，避免了人工提取特征的过程。跌倒行为识别的正确率得到了显著的提升。

1　相关理论与方法

1.1　卷积神经网络

CNN 是一种深度学习网络，最早由 Fukushima[9]在 1980 年提出。通常由输入层、卷积层、池化层、全连接层、输出层构成。卷积神经网络基本结构如图 1。

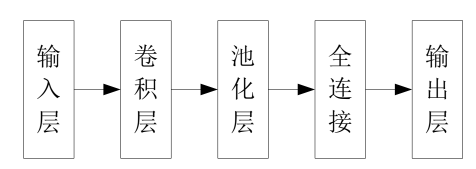


图 1　卷积神经网络结构图

a)输入层。输入层是整个网络的开始，在图像处理领域，卷积神经网络的输入通常为一张图像 X 的像素矩阵。

b)卷积层。卷积层是 CNN 中最重要的一部分。根据对生物视觉细胞局部感受野的理解，卷积层中每一个节点的输入只是上一层神经网络的一小块，卷积层将每一小块进行更加深入的分析从而得到更加抽象的特征。卷积有三种形式，分别是 full、same、valid。以 H_i 表示卷积神经网络第 i 层的特征图（$H_0=X$）。假设 H_i 是卷积层，H_i 的具体产生过程为[10]：

$$H_i=f(H_{i-1}\otimes W_i+b_i) \tag{1}$$

其中：W_i 表示第 i 层卷积核的权值向量；$\otimes$ 表示卷积核与第 $i-1$ 层图像或特征图进行卷积操作；$f(\cdot)$ 表示激活函数，以卷积的输出与第 i 层的偏移量 b_i 代数和作为自变量，通过激活函数 $f(x)$ 得到第 i 层的特征图 H_i。常见的激活函数有 relu、sigmoid、tanh 等。

图 2 以 same 卷积方式为例展示了卷积层的计算过程，其中红色框中为原始矩阵。卷积运算过程是计算两个相同位置元素的乘积之和，图中灰色部分计算过程如下：

（1×0）+（0×0）+（-1×0）+（1×0）+（0×1）+（-1×1）+（1×0）+（0×0）+（-1×1）+1=-1<0。

本例使用 relu 为激活函数，relu 公式为：

$$g(x)=\max(0,x) \tag{2}$$

根据公式（2），最终取值为 0。

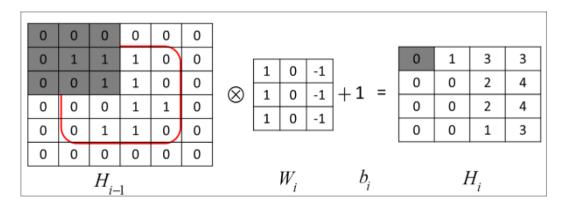


图 2　卷积层计算过程样例图

c)池化层。在卷积层与卷积层之间往往会加上一个池化层（pooling layer），池化层可以非常有效的缩小矩阵尺寸，从而减少最后全连接层中节点的个数，最终达到减少整个神经网络中参数的目的。使用池化层既可以加快计算速度也可以防止过拟合的问题。其中常见的两种池化分别为最大值池化（max-pooling）和平均值池化（average-pooling）。

将图 2 的卷积结果分别进行两种池化操作，池化结果用 R 表示，具体过程如图 3 所示。

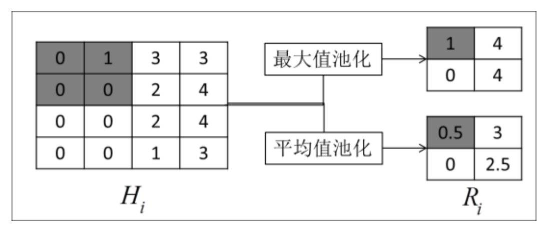


图 3　池化操作样例图

d)全连接层。在卷积神经网络的最后一般会由 1 到 2 个全连接层来给出最后的分类结果。经过几层的卷积层和池化层的处理之后，图像中的信息已经被抽象成了信息含量更高的特征。我们可以将卷积层和池化层看成自动提取图像特征的过程，在特征提取完成以后，仍需要使用全连接层来完成分类的任务。

e)输出层。

常用的输出层为 softmax 层，主要用于分类问题。通过 softmax 层可以得到当前样例属于不同种类的概率分布情况。给定输入 x 属于第 i 类的一种原始度量 $h(x,y_i)$，softmax 公式如下：

$$P(y=i\mid x)=\frac{e^{h(x,y_i)}}{\sum_{j=1}^{n}e^{h(x,y_j)}} \tag{3}$$

其中：$P(y=i\mid x)$ 表示给定输入 x 属于第 i 类的概率。

1.2　长短期记忆网络

长短期记忆网络（Long Short-Term Memory，LSTM）是一种特殊的循环神经网络(RNN)。是为了克服 RNN 网络不能处理远距离依赖的问题而提出的。RNN 中同层隐藏层节点之间有一定的关联，即当序列图片依次输入网络，隐藏层节点的计算不只依赖于当前输入层的输入，也依赖于上一时刻隐藏层各节点的激活值。对于输入序列 $x=(x_1,x_2,\cdots x_t)$，RNN 网络层将得到隐藏层序列 $h=(h_1,h_2,\cdots h_t)$ 和输出序列 $y=(y_1,y_2,\cdots,y_t)$，计算方法如下[6, 11, 12]：

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{4}$$

$$y_t = W_{ho}h_t + b_o \tag{5}$$

其中：H 表示隐藏层所用的激活函数；W_{xh} 表示输入层到隐藏层的权重矩阵；W_{hh} 表示隐藏层到隐藏层的权重矩阵；W_{ho} 表示隐藏层到输出层的权重矩阵；b_h 和 b_o 分别表示隐藏层和输出层的偏向量。

导致 RNN 不能发现序列中时间间隔较长的帧之间的关系的原因是：RNN 没有存储单元来存储和输出信息。不同于标准的 RNN，LSTM 架构[13]使用存储单元来存储和输出信息，从而对较长时间前的输入有了记忆能力。

LSTM 包括新输入 x_t、输出 h_t、输入门 i_t、遗忘门 f_t、输出门 o_t。输入门 i 根据 x_t、c_{t-1}、h_{t-1} 决定哪些部分将进入当前时刻的状态 c_t 进行更新。遗忘门 f_t 决定哪些信息被丢弃。通过遗忘门和输入门，LSTM 结构可以更加有效的决定哪些信息应该被遗忘，哪些信息应该得到保留。具体结构如图 4。

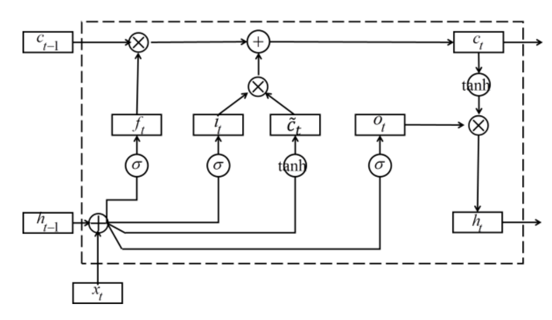


图 4　LSTM 结构图

图 4 中符号⊗表示向量元素乘；符号⊕表示向量拼接；符号⊕表示向量和。LSTM 各组成部分做如下更新[14, 15]：

$$i_t = \sigma(W_{xi}x_t + U_{hi}h_{t-1} + b_i) \tag{6}$$

$$f_t = \sigma(W_{xf}x_t + U_{hf}h_{t-1} + b_f) \tag{7}$$

$$\tilde{c}_t = \tanh(W_{xc}x_t + U_{hc}h_{t-1} + b_c) \tag{8}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{9}$$

$$o_t = \sigma(W_{xo}x_t + U_{ho}h_{t-1} + b_o) \tag{10}$$

$$h_t = o_t \tanh(c_t) \tag{11}$$

其中：σ 表示 sigmoid 激活函数；⊙表示向量元素乘；W_{xi}、W_{xf}、W_{xc}、W_{xo} 分别表示输入层到输入门、遗忘门、存储单元 cell 和输出门之间的权重矩阵；U_{hi}、U_{hf}、U_{hc}、U_{ho} 分别表示隐藏层到输入门、遗忘门、存储单元 cell 和输出门之间的权重矩阵；b_i、b_f、b_c、b_o 分别表示输入门、遗忘门以及存储单元 cell 和输出门的偏置值；i、f、o、c 分别表示输入门、遗忘门、输出门和存储单元。

2　跌倒行为识别的混合深度神经网络模型

2.1　基本思想

将视频数据经过预处理所得序列图片随机分为训练数据集和测试数据集。训练数据用于模型的构建和参数的调整，测试数据用于检验模型的性能。利用 CNN 网络提取各帧的空间特征，然后将 CNN 网络的输出调整规模依次输入到 LSTM 网络来获取序列时序特征，并计算各个时刻 LSTM 输出的平均值，预测最后的分类结果。混合模型基本结构如图 5 所示。

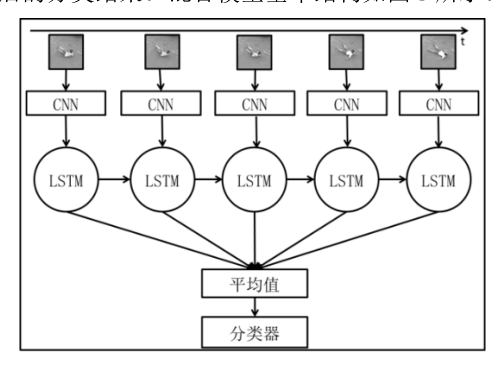


图 5　混合模型结构

2.2　混合深度神经网络模型

2.2.1 卷积神经网络处理层

模型使用卷积神经网络来提取视频帧的空间信息，生成行为的表示特征。

令 N 表示输入网络的图像序列的帧数。对于单张图像，像素矩阵的大小为 $P \times Q$，采用 same 方式卷积，令卷积核大小为 $k \times k$，需要在原始图像矩阵的外围加上长度为 $k/2$ 的零填充为 $(P+2\,k/2) \times (Q+2\,k/2)$ 大小的像素矩阵，其中：　表示向下取整符号。令 v_{ij} 表示扩充后像素矩阵中 (i, j) 处像素值。则落入第 $(Q(i-1)+j)$ 个 $(i \le P+2\,k/2\ , j \le Q+2\,k/2)$ 滑动窗口的所有像素值可以表示为窗口矩阵，如下所示：

$$X_{ij} = \begin{pmatrix} v_{ij} & \cdots & v_{i(j+k-1)} \\ \vdots & \ddots & \vdots \\ v_{(i+k-1)j} & \cdots & v_{(i+k-1)(j+k-1)} \end{pmatrix} \tag{12}$$

对于每个窗口矩阵，结合公式（1）进行卷积运算得到当前窗口特征。

$$Y_{ij} = f(x_{ij} \otimes W + b) \tag{13}$$

在卷积运算过程中，鉴于 relu 收敛速度快的特性，f 采用如下 relu 激活函数：

$$g(x) = \max(0, x) \tag{14}$$

在完成卷积之后进行池化操作。选择最大池化来进行处理。获得每一个窗口矩阵的最大特征值。

$$R_n = Max(Y_{ij}) \tag{15}$$

其中：R_n 表示序列图像中第 n 张图像经过卷积和池化操作以后的特征矩阵。

分别对序列图像进行以上操作，则对于序列中各个图像帧的特征矩阵可用 $R = (R_1, R_2, \cdots R_n)$ 表示，其中 $n \le N$。

2.2.2 长短期记忆模型处理层

CNN 层的一个输出 R 对应一个时刻 t 的 LSTM 输入。某时刻 t，根据公式（6）~（11），LSTM 单元各组成部分做如下更新：

$$i_t = \sigma(W_{ri}R_t + U_{hi}h_{t-1} + b_i) \tag{16}$$

$$f_t = \sigma(W_{rf}R_t + U_{hf}h_{t-1} + b_f) \tag{17}$$

$$\tilde{c}_t = \tanh(W_{ri}R_t + U_{hc}h_{t-1} + b_c) \tag{18}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{19}$$

$$o_t = \sigma(W_{ro}R_t + U_{ho}h_{t-1} + b_o) \tag{20}$$

$$h_t = o_t \tanh(c_t) \tag{21}$$

其中：σ 表示 sigmoid 激活函数；R_t 表示 t 时刻输入的特征矩阵；W_{ri}、W_{rf}、W_{rc}、W_{ro} 分别表示输入层到输入门、遗忘门、存储单元 cell 和输出门之间的权重矩阵；U_{hi}、U_{hf}、U_{hc}、U_{ho} 分别表示隐藏层到输入门、遗忘门、存储单元 cell 和输出门之间的权重矩阵；b_i、b_f、b_c、b_o 分别表示输入门、遗忘门、存储单元 cell 和输出门的偏置值。

3　实验测试与分析

3.1　数据集与测试环境

论文使用 CASIA 数据集[16]作为测试数据，该数据集是由中国科学院自动化研究所提供。所有视频都是由分布在水平视角、斜角和俯角的三个未标定的静止的摄像机同时拍摄的，帧率为 25fps，采用 huffyuv 编码压缩，分辨率为 320*240。图 6 为选取的原始视频帧。



图 6　视频帧序列

利用基于 Python 的深度学习库 Keras 在 GPU 加速环境下进行实验。具体实验环境如表 1 所示。

表 1　实验环境配置

实验环境	配置
操作系统	Ubuntu14.04
GPU	NVIDIA　Tian XP
内存\硬盘	64GB\4.2TB
程序语言	Python3.6
程序框架	Keras

3.2　实验方案与参数设置

选择 CASIA 数据集中俯视角下单人行为的弯腰走、下蹲、晕倒、跳、跑、走作为实验数据。其中跌倒为异常数据集，弯腰走、下蹲、跳、跑、走作为正常数据集。经过对视频数据预处理后获得跌倒序列图片 955 张，非跌倒序列图片 840 张。其中随机选择 80%作为训练数据集，20%作为测试数据集。

由于原始数据为彩色图像序列，色彩通道存在不稳定特性，所以将原始图像进行预处理，转化为单通道图像，并将像素值简单缩放归一化到[0,1]区间，用于最终的实验数据。

为了获取实验输入序列最佳帧数并验证混合模型的有效性，本文采用在相同的实验环境下，相同数据集以及相同数据量，分别做如下两组对比试验：

a)分别采取序列帧数为 3、4、5、6、7 输入混合模型进行实验。

b)对 SVM、CNN[4]、LSTM[5]以及本文采用的混合模型进行跌倒检测的对比实验。

深度学习模型主要涉及的参数有：滑动窗口大小、卷积核数、激活函数、LSTM 节点数、优化方法以及学习率。选择优化方法（选取：Adam、SGD、RMSprop）、学习率（取值：0.0001，0.001，0.01，0.1），LSTM 节点数（选取：32,64,128）和激活函数（选取：relu，tanh）进行实验取优，其余参数为默认值，其中损失函数采用交叉熵损失函数，通过实验对比可知，模型中参数分别设置如表 2 所示，模型性能达到最佳。

表 2　参数设置

参数名称	参数
学习率	0.0001
优化函数	SGD
激活函数	relu
损失函数	categorical_crossentropy
LSTM 节点数	64

3.3　实验结果及分析

通过对不同序列帧数进行对比实验，结果如图 7 所示。

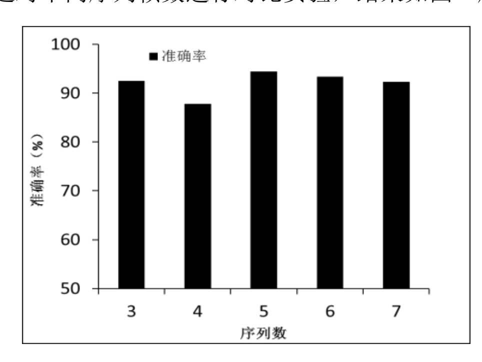


图 7　不同帧数实验对比结果

由图 7 所知，当序列数为 5 时实验效果达到最佳，这可能是由于序列帧数太小会丢失部分跌倒的行为信息，不能很好的表示跌倒行为，而序列帧数过大则导致总体训练样本数变小，不能很好的训练模型。所以实验采用每 5 帧为一个序列进行实验，并用准确率对模型进行评价，准确率实验结果如表 3 所示。

表 3　各个模型识别准确率

模型	准确率
SVM	82.17%
CNN	83.84%
LSTM	91.67%
CNN+LSTM	94.44%

另外，对深度学习模型在 GPU 加速环境下训练时间也进行

了对比实验，结果如图 8 所示。

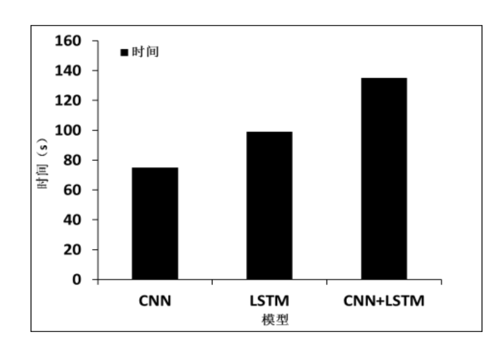


图 8　各模型训练时间对比

从表 3 可以看出，对 CASIA 数据库进行跌倒检测的准确率按由高到低的顺序依次为：CNN+LSTM、LSTM、CNN、SVM。CNN+LSTM 混合模型的准确率要高于其它三种模型，这得益于混合模型既有效利用了 CNN 通过卷积获取局部空间特征，又结合了 LSTM 的时序性来获得视频序列的时间特征。另外，通过深度学习的方法也避免了繁杂的人工提取特征，泛化能力更强。但同时由图 8 可知，在模型的训练时间上，CNN 和 LSTM 混合模型由于复杂的网络结构，其耗时分别高于 CNN 和 LSTM 的耗时，而 LSTM 耗时高于 CNN 是因其带有记忆功能，网络结构更为复杂。

4　结束语

论文提出了基于 CNN 和 LSTM 混合模型来检测跌倒行为，在 CASIA 数据集上的识别率达到了 94.44%。相比较浅层传统模式识别方法避免了人工提取特征，增强了模型的泛化能力；对于深层 CNN 和 LSTM 网络不但能够提取到序列视频帧的空间特征，也能提取到帧与帧之间的时序性信息，识别率提升明显，具有一定的可靠性。由于实验中采用的数据集背景固定单一，且都为单人行为，与实际情况还有偏差。未来应对更加接近于实际场景中的跌倒行为进行深入的研究和分析。

参考文献：

[1] Mubashir M, Shao L, Seed L. A survey on fall detection: principles and approaches [J]. Neurocomputing, 2013, 100 (2): 144-152.

[2] 汪大峰，刘勇奎，刘爽，等. 视频监控中跌倒行为识别 [J]. 电子设计工程, 2016, 24 (22): 1126-1222. (Wang Dafeng, Liu Yongkui, Liu Shuang, etal. Abnormal behavior recognition of fall in surveillance video [J]. Electronic Design Engineering, 2016, 24 (22): 1126-1222.)

[3] 白勇，孙晓雯，秦昉，等. 基于 SVD 特征降维和支持向量机的跌倒检测算法 [J]. 计算机应用与软件，2017, 34 (1): 247-251. (Bai Yong, Sun Xiaowen, Qin fang, etal. The falling detection algorithm based on SVD feature dimension reduction and SVM [J]. Computer Applications and Software. 2017, 34 (1): 247-251)

[4] 王忠民，曹洪江，范琳. 一种基于卷积神经网络深度学习的人体行为识别方法 [J]. 计算机科学，2016, 43 (s2): 56-58. (Wang Zhongmin, Cao Hongjiang, Fan Lin, etal. Method on human activity recognition based on convolutional neural networks [J]. Computer Science, 2016, 43 (s2): 56-58.)

[5] 匡晓华，何军，胡昭华，等. 面向人体行为识别的深度特征学习方法比较 [J]. 计算机应用研究, 2018, 35 (9): 2815-2817, 2822. (Kuang Xiaohua, He Jun, Hu Shaohua, *et al.* Comparison of deep feature learning methods for human activity recognition [J]. Application Research of Computers, 2018, 35 (9): 2815-2817, 2822)

[6] Ng Y H, Hausknecht M, Vijayanasimhan S, *et al.* Beyond short snippets: deep networks for video classification [C]// Computer Vision and Pattern Recognition. 2015: 4694-4702.

[7] Venugopalan S, Xu H, Donahue J, *et al.* Translating Videos to Natural Language Using Deep Recurrent Neural Networks [J]. Computer Science, 2015.

[8] Ma Xuezhe, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf [EB/OL]. (2016-05-29) . https://arxiv. org/abs/1603. 01354.

[9] Fukushima K. Neocognitron: A hierarchical neural network capable of visual pattern recognition [J]. Neural Networks, 1988, 1 (2): 119-130.

[10] 李彦冬，郝宗波，雷航. 卷积神经网络研究综述 [J]. 计算机应用，2016, 36 (9): 2508-2515. (Li Yandong, Hao Zongbo, Lei Hang. Survey of convolutional neural network [J]. Journal of Computer Applications, 2016, 36 (9): 2508-2515.)

[11] Graves A. Supervised sequence labelling with recurrent neural networks [M]. Springer Berlin Heidelberg, 2012.

[12] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 6645-6649.

[13] Gers F A, Schraudolph N N. Learning precise timing with lstm recurrent networks [M]. JMLR. org, 2003.

[14] Sundermeyer M, Ney H, Schlyuter R. From feedforward to recurrent LSTM neural networks for language modeling [J]. IEEE//ACM Trans on Audio, Speech & Language Processing, 2015, 23 (3): 517-529.

[15] Cai M, Liu J. Maxout neurons for deep convolutional and LSTM neural networks in speech recognition [J]. Speech Communication. 2016, 77 (C): 53-64.

[16] 中国科学院行为分析数据库 [EB/OL]. http://www. cbsr. ia. ac. cn/china/Action Databases CH. asp. (Behavioral analysis database of Chinese Academy of Sciences [EB/OL]. http://www. cbsr. ia. ac. cn/china/Action Databases CH. asp.)